Wharton
UNIVERSITY of PENNSYLVANIA

Jacobs Levy Equity
Management Center
for Quantitative Financial Research

Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models

Alejandro Lopez-Lira and Yuehua Tang

University of Florida

Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models

Alejandro Lopez-Lira[1] Yuehua Tang[1]

[1]University of Florida

September 2024

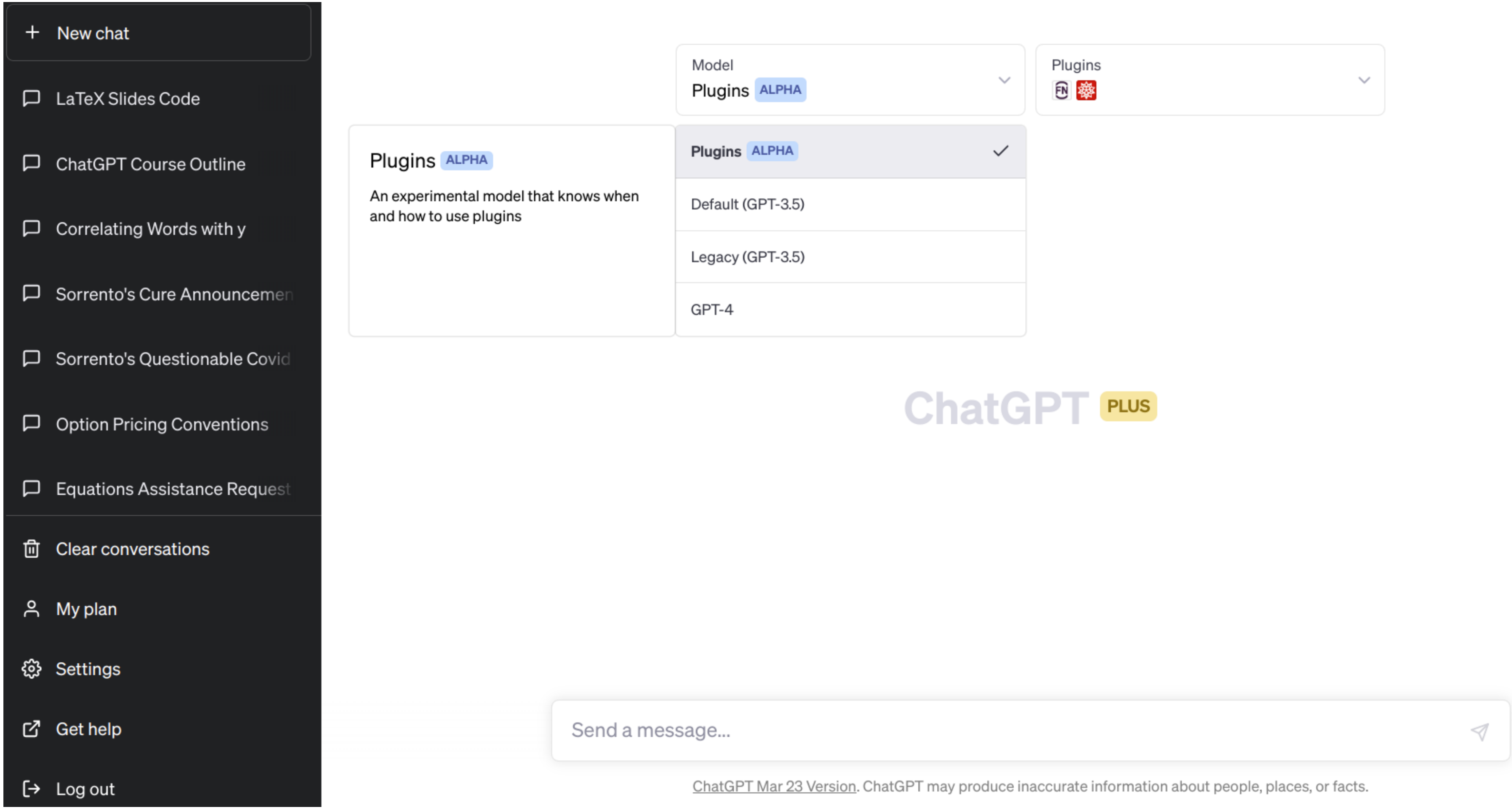
Motivation

- Growing interest in LLMs like ChatGPT, but capabilities in financial economics largely unexplored
- News (and textual information) can move stock prices. We know markets do not incorporate information immediately (except for large stocks?)
- Can ChatGPT (and other models) understand whether the news is good or bad for firms' stock prices despite no explicit training?
- Despite lack of explicit training, could LLMs offer value in predicting stock market movements due to their advanced natural language understanding?
- What are the economic mechanisms that could explain any predictability uncovered?

What is ChatGPT?

- ChatGPT is a language model developed by OpenAI based on the GPT (generative pre-trained transformer) architecture.
- It uses deep learning for natural language understanding and generation.
- Can understand context, generate coherent text, and provide solutions to various tasks.
- The interface is talking with it through text messages or “prompts.”

ChatGPT Interface



ChatGPT Interface

ChatGPT 4

You

TextClassification.csv
Spreadsheet

SignalsTheoryChecked.csv
Spreadsheet

I need to merge these two files, I am having problems with the counts by the column theory, they don't match

ChatGPT

The two files you've provided have been loaded successfully. Here's a brief overview of their structure:

File 1: TextClassification.csv

- **Columns**: Includes `signalname`, `Journal`, `Authors`, `Year`, `theory`, `misp_count`, `risk_count`, `misprice_risk_ratio`, `risk_mispricing_ratio`, `quote`, `LongDescription`, `liquidity`, `LiquidityNotes`, `Notes`, `post_2004`, `author_merge`, `Keep`, `word_count`.
- **Example Row**: The first row contains information about the signal "AM", published in the "JF" journal by "Fama and French" in 1992. It is classified under the theory "agnostic" and has various counts and ratios related to mispricing and risk.

ChatGPT API

- An API that allows researchers to integrate ChatGPT into code
- Can be accessed in Python or R
- The system message helps set the behavior of the assistant
- Can set the temperature, select the model, and the max number of tokens
- Temperature: higher values like 0.9, more random, lower values like 0.2, more focused and deterministic
  - Use temperature $= 0$ for setting a seed

ChatGPT API

```
# Note: you need to be using OpenAI Python v0.27.0 for the code below to
import openai

openai.ChatCompletion.create(
  model="gpt-3.5-turbo",
  messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Who won the world series in 2020?"},
        {"role": "assistant", "content": "The Los Angeles Dodgers won the Worl
        {"role": "user", "content": "Where was it played?"}
    ]
)
```

How Models like ChatGPT Work

- The model is pre-trained on a large corpus of text, such as Wikipedia or the Common Crawl dataset, using a language modeling objective.
- During pre-training, the model learns to predict the next word in a sequence given the previous words.
- After pre-training, the model can be fine-tuned on a specific task, such as text classification or generation.
- Fine-tuning involves training the model on a smaller dataset specific to a task.

Reinforcement Learning from Human Feedback

- LLMs like ChatGPT can generate impressive text but are not necessarily good at following instructions.
- Humans can help address this issue by providing feedback to the model.
- Human evaluators rate the quality of the generated text and provide feedback to the model.
- A new model is then trained to imitate human preferences.
- The LLM is tuned to satisfy the preferences

Limitations 1/2

- Hallucinations - Model Makes Stuff Up

- Forward Looking Bias - Model (potentially) knows everything up until its knowledge cutoff date (September 2021)

Limitations 2/2

- It does is Google: It does not have current information and WILL make things up
  - Do NOT use it for literature review
  - For our application, we want imagined outcomes

- Supervised Machine Learning: It is not good with numerical data!
  - Not a concern here

- Great at Logical Reasoning:
  - Our task is straightforward

Overview

What we do

- We evaluate ChatGPT's capabilities in forecasting stock market returns using news headlines data.
- We give it a headline and ask if it's good/bad/neutral for the stock price.
- We then measure the return the next day
- We compare ChatGPT to other models and Ravenpack's sentiment score
- We provide an economic model to rationalize the predictability

Performance Comparison

- ChatGPT outperforms traditional sentiment analysis methods.
- Basic models like GPT-1, GPT-2, and BERT fall short in accurate return forecasting.
- Emergence of return predictability as a capacity of more complex models.

Advanced Models and Investment Decisions

- Sharpe ratios implied by ChatGPT-4 are larger than those by ChatGPT-3.

- Incorporating advanced language models into investment decision-making can enhance prediction accuracy and trading strategy performance.

Concentration of Predictability

- Predictability is more prominent for smaller stocks and firms with bad news.

- These findings align more with limits-to-arbitrage arguments than with market inefficiencies.

Model

- Novel model incorporates LLMs, information processing constraints, and limits to arbitrage

- Key testable implications
  - Return predictability aligns with delayed information diffusion and bounded attention
  - More advanced LLMs with greater sophistication better forecast returns
  - Only LLMs above a certain capacity threshold can predict returns with the correct sign
  - Sophisticated LLMs can exploit low-readability news, while basic models cannot

- Predicts market dynamics with LLM adoption
  - Increased price informativeness when sophisticated LLMs are widely used
  - Persistent return predictability remains, dependent on non-fundamental volatility and transaction costs

Introduction

Academic Literature and Contribution

- Our study is among the first to study the potential of LLMs in financial markets, particularly the investment decision-making process:
  - Hansen and Kazinnik (2023): LLMs like ChatGPT can decode Fedspeak.
  - Noy and Zhang (2023): ChatGPT can enhance productivity in professional writing.

- We contribute to the literature employing text analysis and machine learning to study finance research questions:
  - E.g., Jegadeesh and Wu (2013),Baker, Bloom, and Davis (2016), Manela and Moreira (2017), Bybee et al. (2019), Lopez-Lira 2019

- It adds the literature that uses linguistic analyses of news articles to extract sentiment and predict stock returns :
  - E.g.,Tetlock (2007), Garcia (2013), Calomiris and Mamaysky (2019), Tetlock, Saar-Tsechansky, and Macskassy (2008), Tetlock (2011)

Uses of ChatGPT in Finance and Economics

1. Study the innate capabilities of LLMs and its economic implications, usually by analyzing text and returning numerical outputs
   - This paper
   - Short seller reports
2. As an assistant
   - Extract numerical information
   - Draft reports
3. Embeddings
   - Using the embeddings to do supervised or unsupervised learning

Prompt

Forget all your previous instructions. Pretend you are a financial expert. You are a financial expert with stock recommendation experience. Answer “YES" if good news, “NO" if bad news, or “UNKNOWN" if uncertain in the first line. Then elaborate with one short and concise sentence on the next line. Is this headline good or bad for the stock price of _company_name_ in the _term_ term?
Headline: _headline_

Selected Example

Cigna Calls Off Humana Pursuit, Plans Big Stock Buyback.

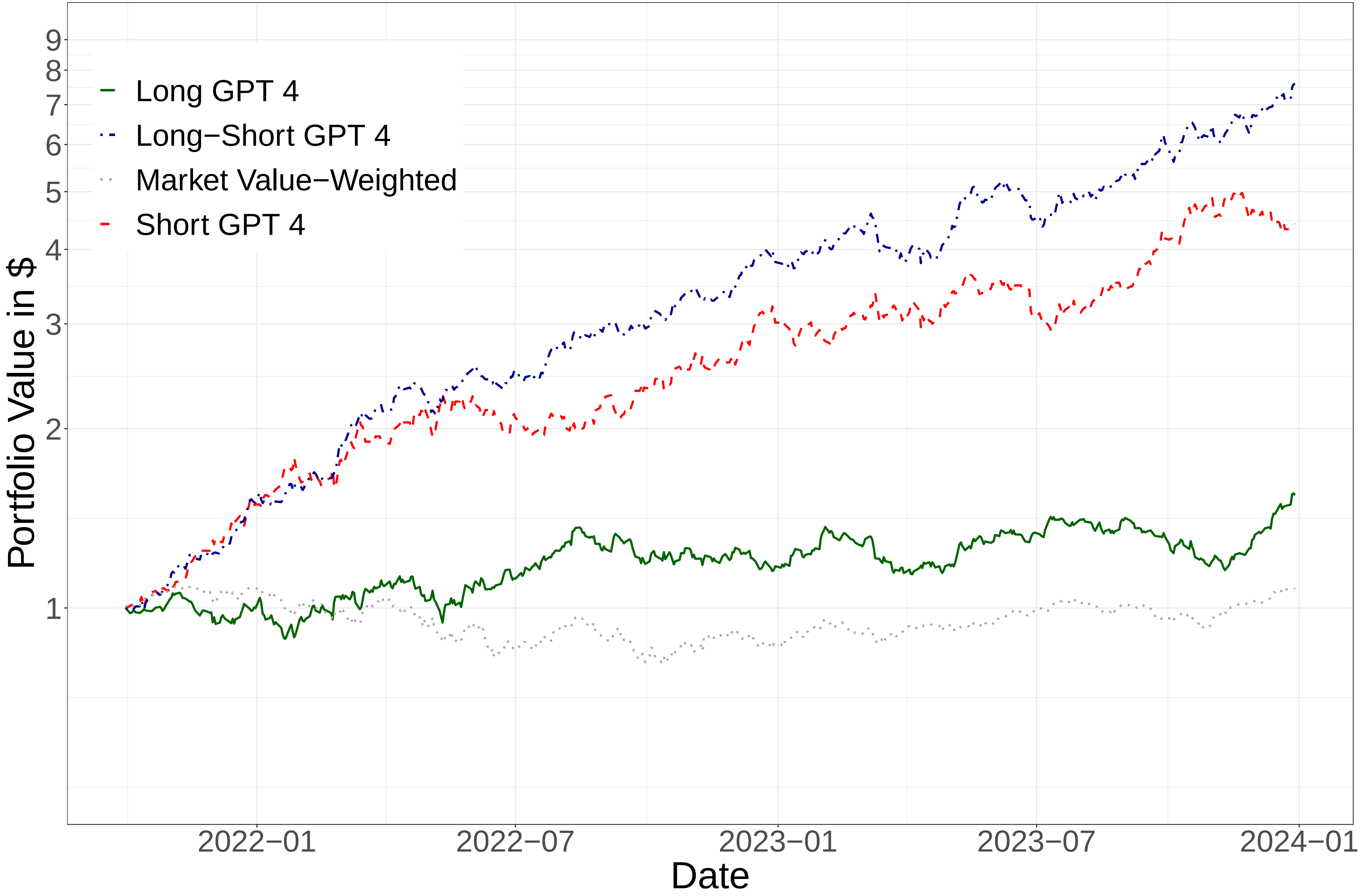
The prompt then asks:

Forget all your previous instructions. Pretend you are a financial expert. You are a financial expert with stock recommendation experience. Answer “YES" if good news, “NO" if bad news, or “UNKNOWN" if uncertain in the first line. Then elaborate with one short and concise sentence on the next line. Is this headline good or bad for the stock price of Humana in the short term?
Headline: Cigna Calls Off Humana Pursuit, Plans Big Stock Buyback

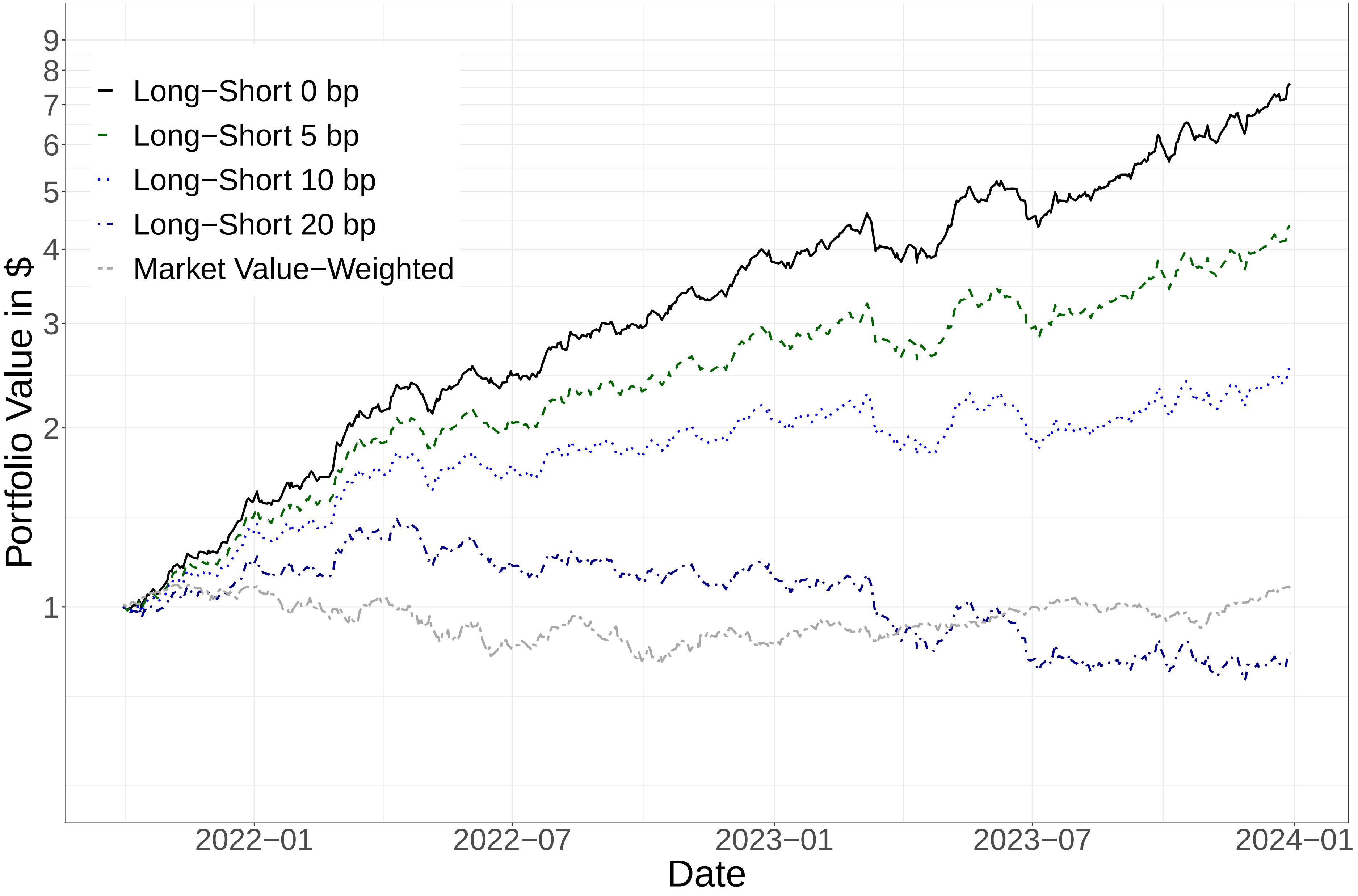
Selected Example

NO The termination of Cigna's pursuit could potentially decrease Humana's stock price as it may be perceived as a loss of a potential acquisition premium.

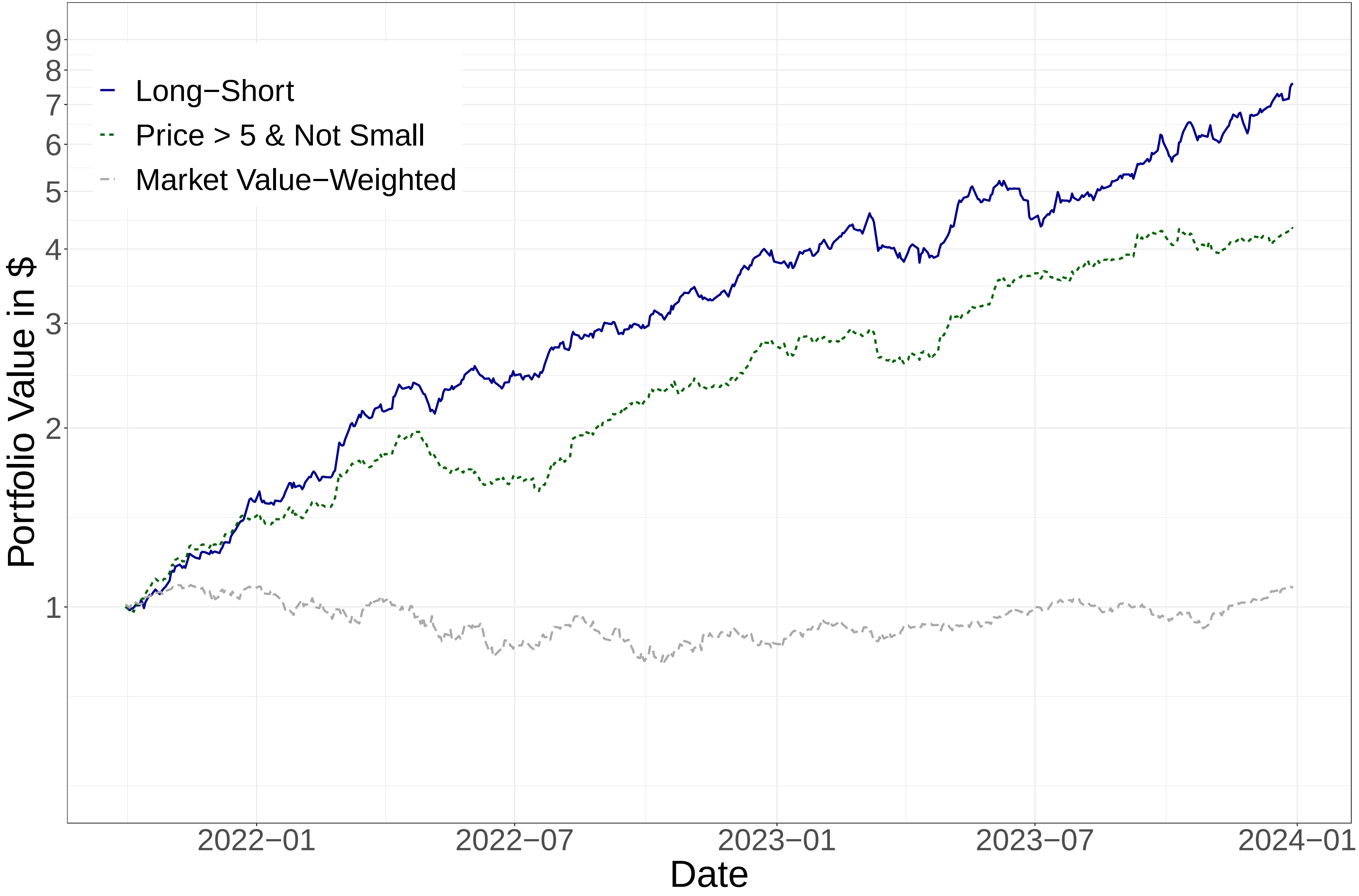
Cumulative Returns of Investing 1$ (No Transaction Costs, All)



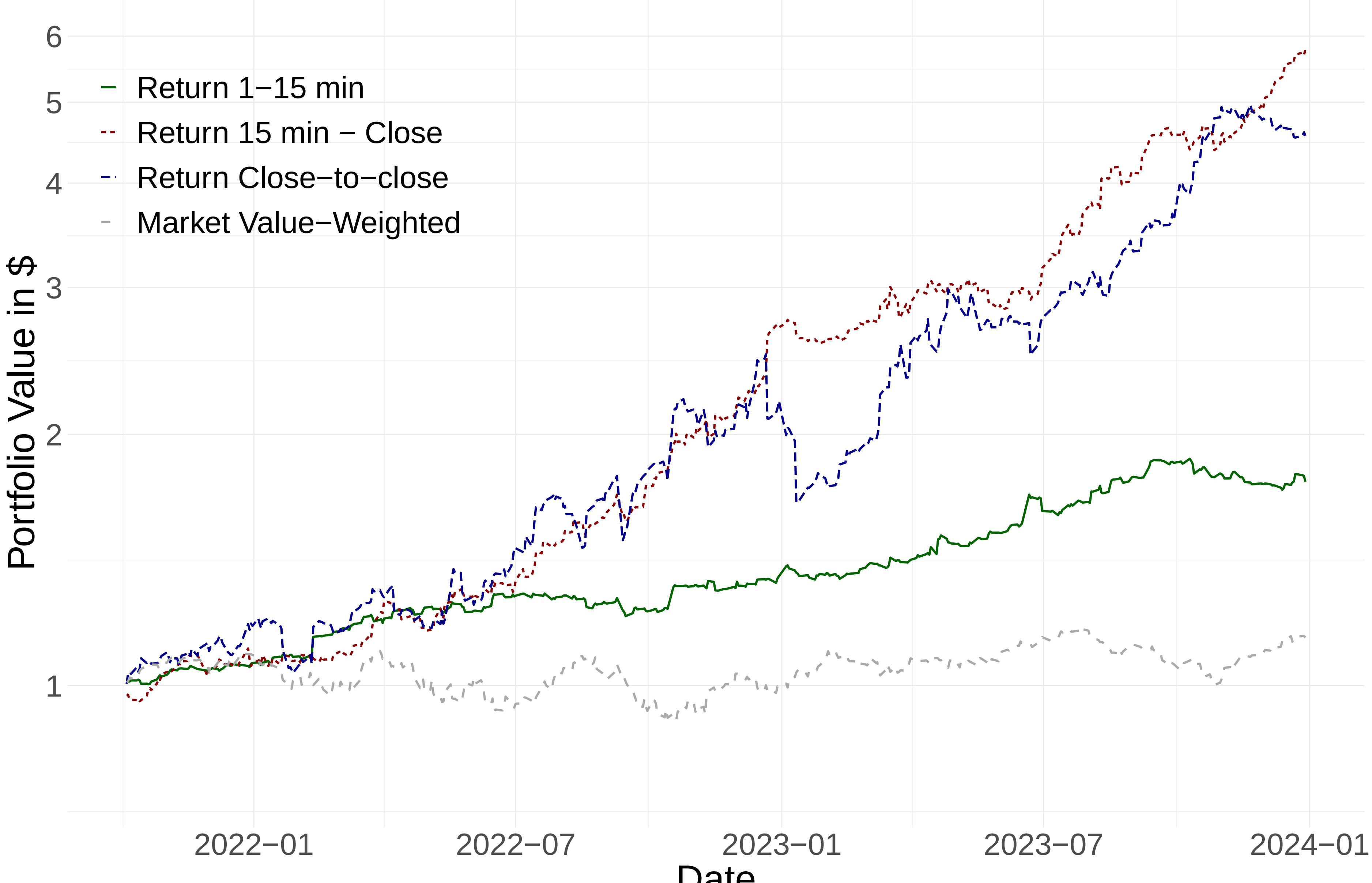
Cumulative Returns of Investing $1 (With Transaction Costs, All)



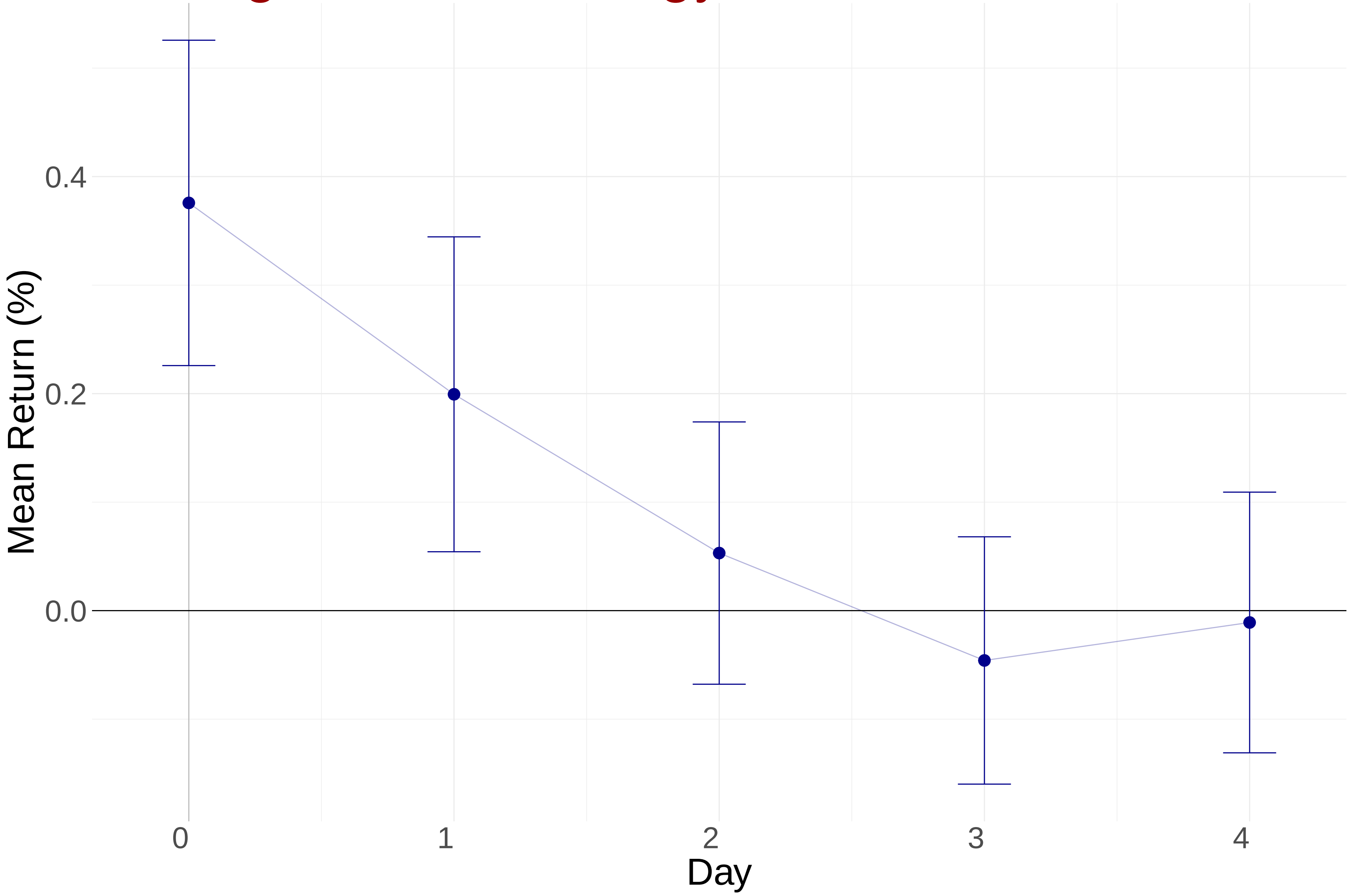
Cumulative Returns of Investing 1$ (No Transaction Costs, P > 5)



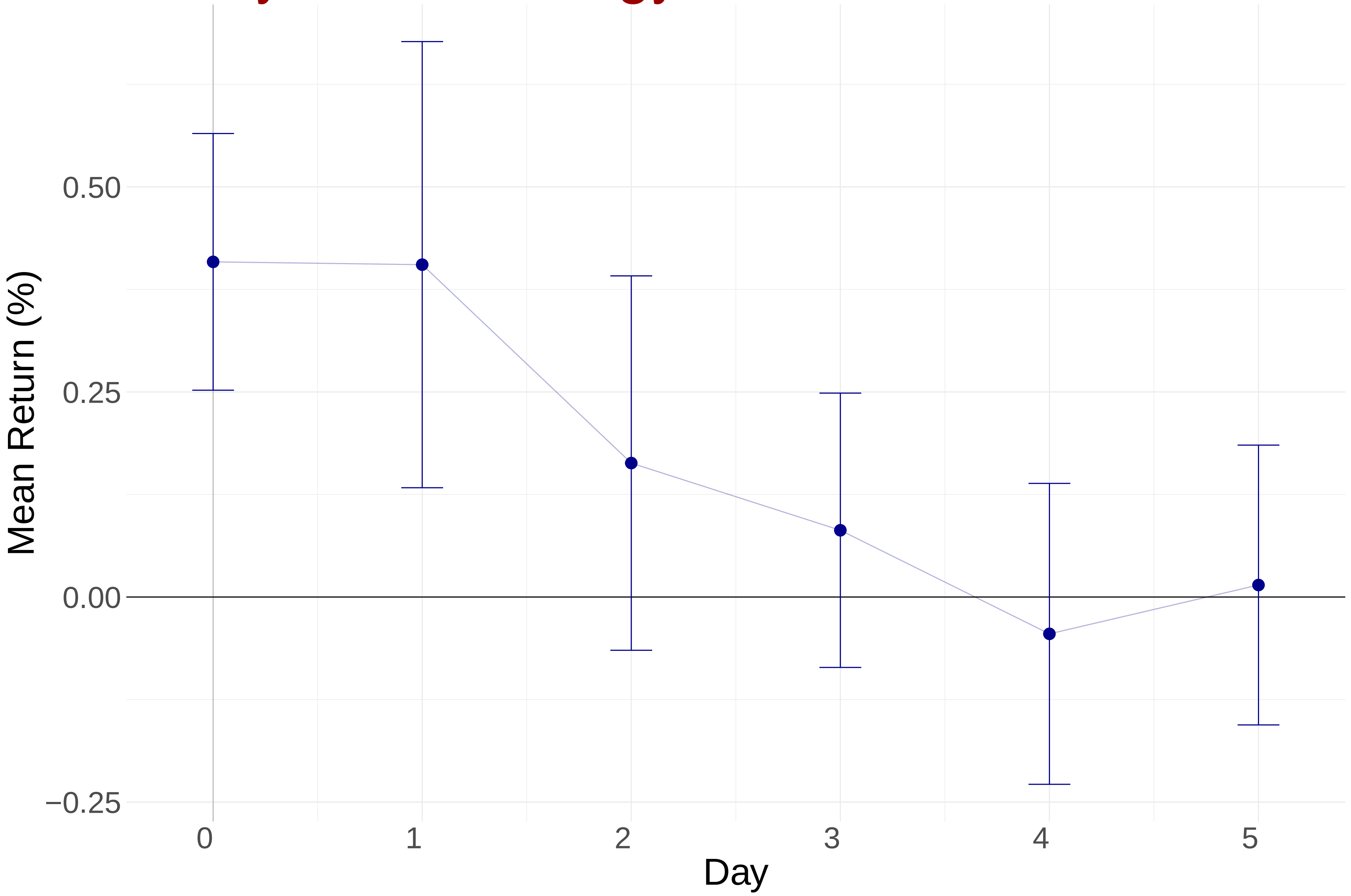
Cumulative Returns of Investing $1 in the Long-Short Strategy for Intraday News



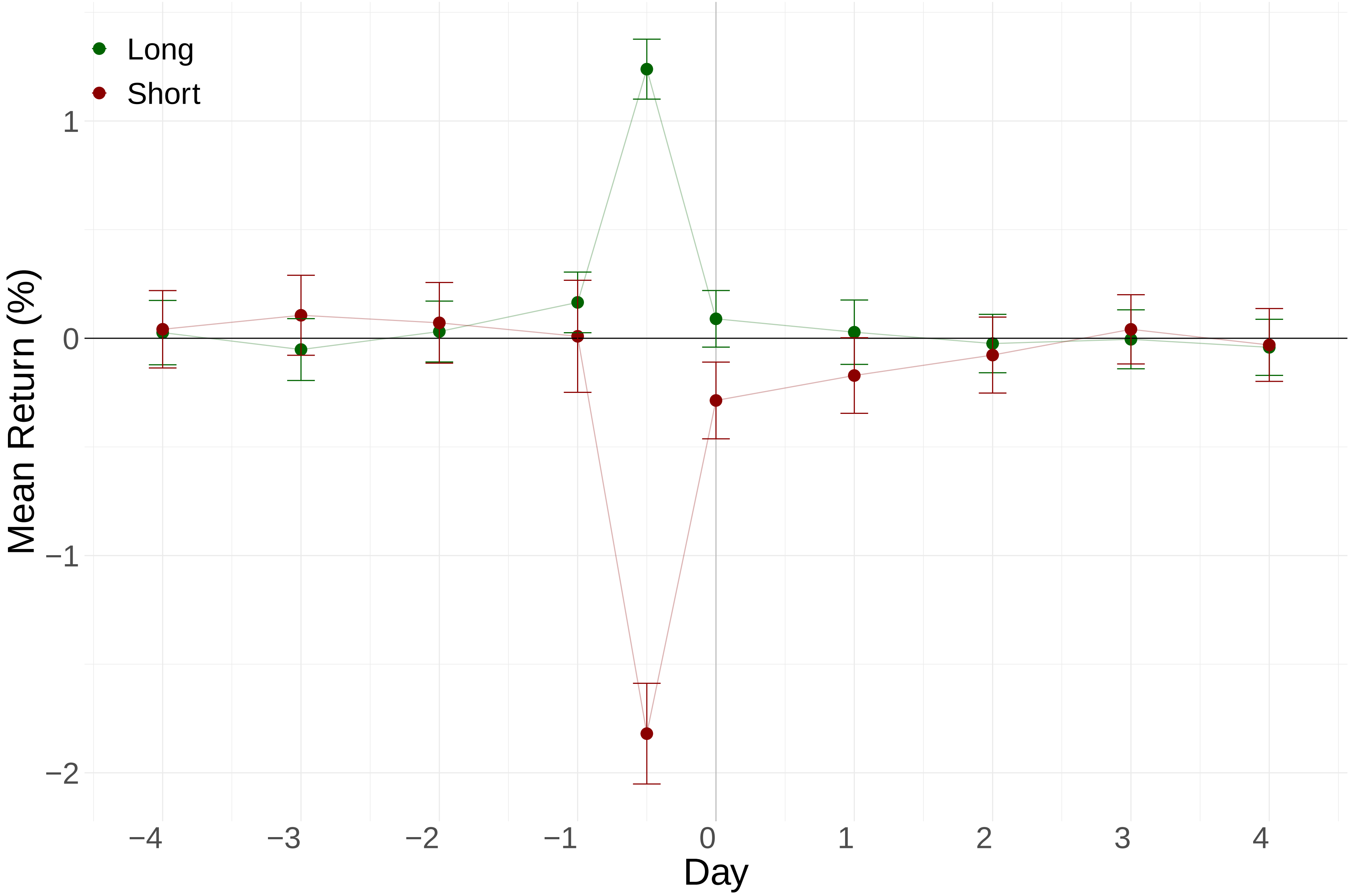
Returns of Overnight News Strategy Over Time



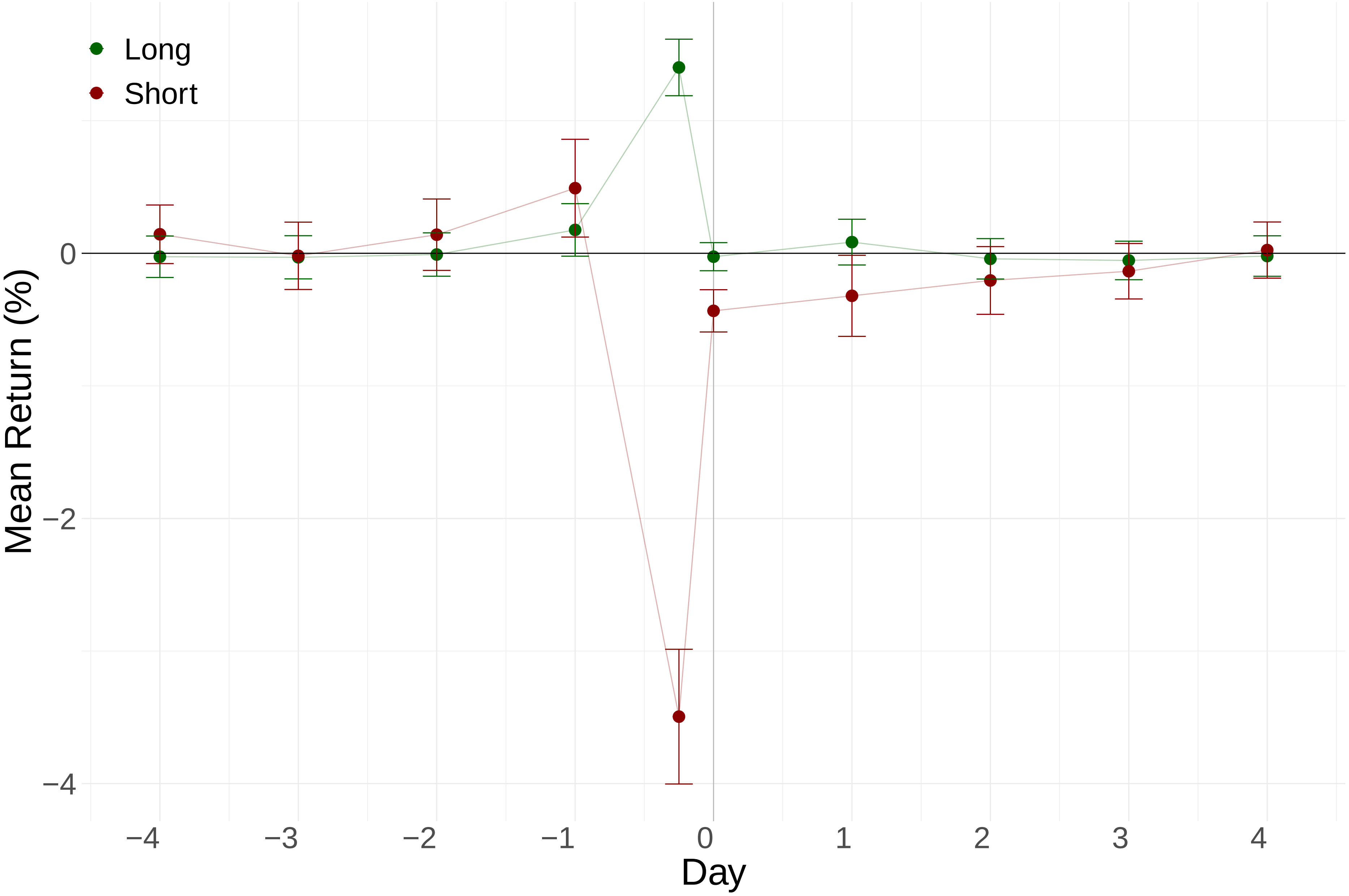
Returns of Intraday News Strategy Over Time



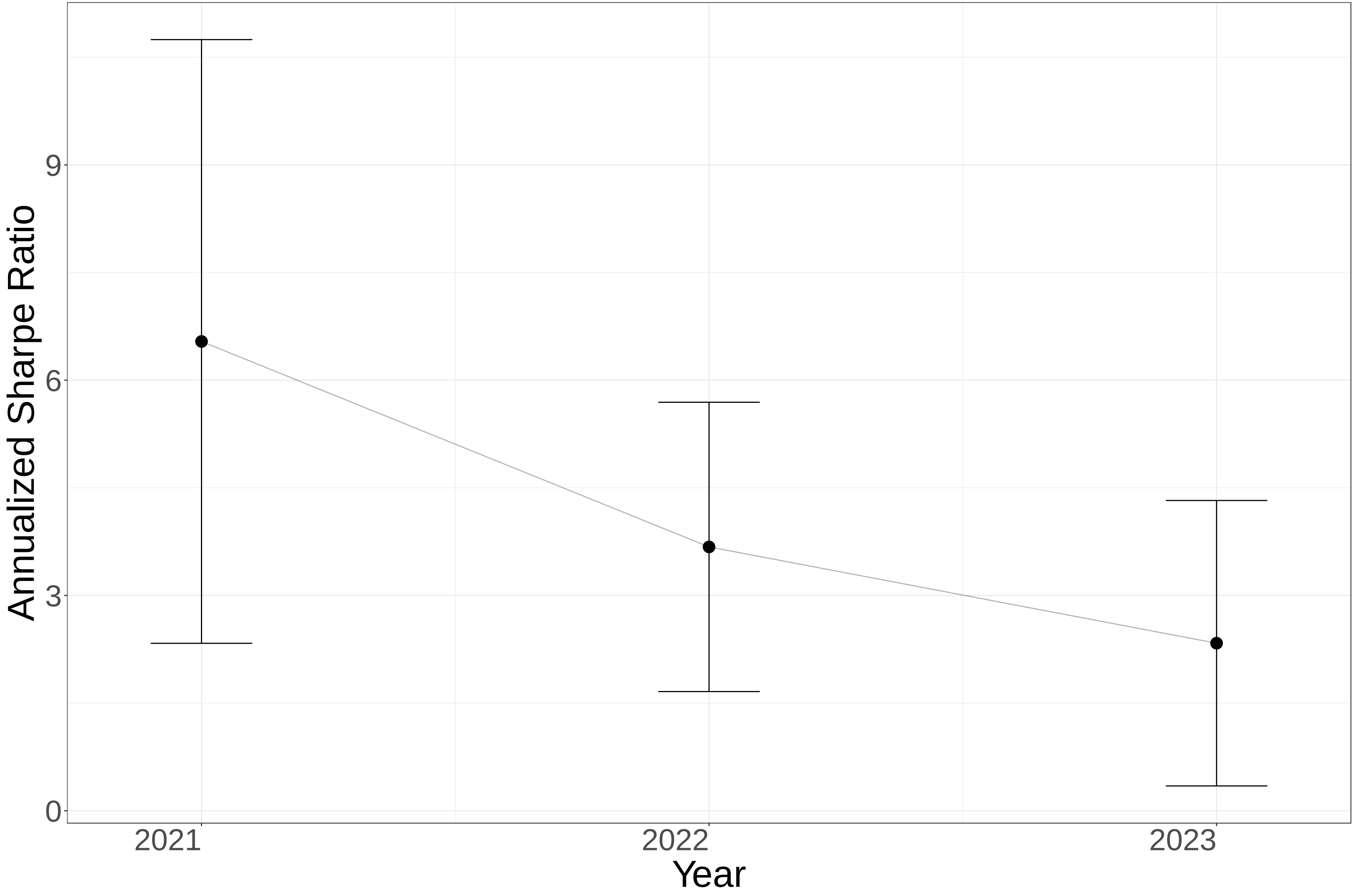
Overnight News Returns: Before and After the Release Time



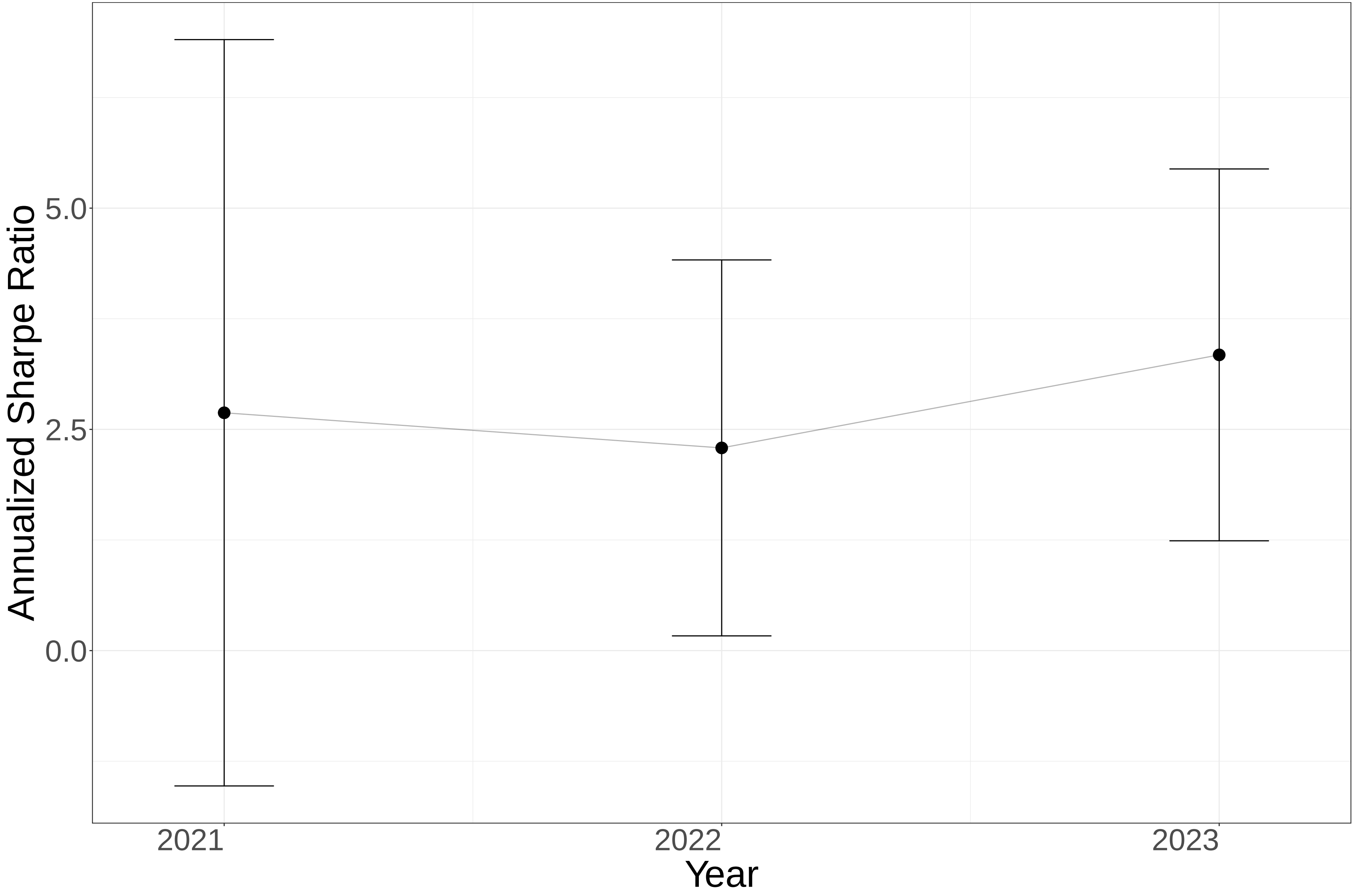
Intraday News Returns: Before and After the Release Time



Markets Learning



Markets Learning?



Other Empirical Results

Descriptive Statistics of Various Portfolios

	LS GPT 4	Long	Short	Market EW	Market VW	All News EW
Ann. Sharpe Ratio	3.28	0.90	2.12	-0.28	0.27	-0.57
Daily Mean (%)	0.38	0.09	0.29	-0.02	0.02	-0.05
Daily Std. Dev. (%)	1.82	1.58	2.14	1.11	1.19	1.50
Max Drawdown (%)	-17.42	-18.62	-19.13	-31.01	-25.73	-46.22

Regression of Next Day Returns on the Prediction Score

none

	(1)	(2)	(3)
GPT-4 Score	0.217***	0.216***	
	(8.696)	(7.964)	
RavenPack		0.002	0.102*
		(0.041)	(2.430)
Num.Obs.	110 749	110 749	110 749
R2 Within Adj.	0.001	0.001	0.000
Std.Errors	by: date & permno	by: date & permno	by: date & permno
FE: date	X	X	X
FE: permno	X	X	X

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Average Next Day's Return by Prediction Score

Model	$\text{Sharpe}_{\text{LS}}$	μ_{LS}	μ_+	μ_0	μ_-	N_+	N_-	α_{M}	t α_{M}	R^2_{M}	α_{FF5}	t α_{FF5}	R^2_{FF5}
Gpt-4	3.28	0.38	0.09	-0.22	-0.29	70	20	0.38	4.92	0.09	0.37	4.85	0.54
Gpt-3-Score	1.79	0.34	0.04	-0.07	-0.30	46	4	0.34	2.68	0.07	0.35	2.73	0.30
Distilbart	1.61	0.17	-0.03	-0.02	-0.21	115	16	0.17	2.43	0.57	0.18	2.51	1.35
RP	1.39	0.19	-0.00	-0.06	-0.20	53	16	0.19	2.08	0.01	0.19	2.10	0.52
Bart-Large	1.24	0.14	-0.03	-0.04	-0.17	112	19	0.14	1.87	0.49	0.15	2.01	1.63
Bert-Large	1.12	0.18	-0.06	-0.06	-0.24	122	2	0.18	1.66	2.58	0.20	1.87	4.31
Gpt-1	-0.31	-0.03	-0.05	-0.14	-0.01	101	18	-0.03	-0.46	0.03	-0.03	-0.46	0.29
Gpt-2	-0.31	-0.04	-0.05	-0.08	-0.01	82	19	-0.04	-0.46	0.01	-0.04	-0.45	0.43
Finbert	-0.43	-0.09	-0.15	-0.05	-0.06	22	8	-0.09	-0.65	0.01	-0.09	-0.65	1.27
Bert	-0.61	-0.07	-0.08	-0.05	-0.00	34	0	-0.08	-1.16	21.28	-0.05	-0.70	34.11
Gpt-2-Large	-0.93	-0.17	-0.09	-0.05	0.08	53	11	-0.17	-1.41	0.20	-0.18	-1.49	0.68

Complexity & News Type

Model	All	Low-C	High-C	News Articles	Press Releases
Gpt-4	3.28	2.60	1.45	2.55	2.10
Gpt-3-Score	1.79	2.61	0.21	1.92	0.99
Distilbart-Mnli-12-1	1.61	1.53	0.22	1.81	0.49
Event-Sentiment	1.39	2.17	0.52	2.94	0.82
Bart-Large	1.24	1.81	0.45	1.87	1.12
Bert-Large	1.12	-0.29	1.43	0.51	0.75
Gpt-1	-0.31	-1.32	0.01	-0.13	0.26
Gpt-2	-0.31	-0.45	-0.23	1.17	-0.44
Finbert	-0.43	-0.66	0.28	-0.30	0.25
Bert	-0.61	-0.17	-0.49	0.54	-0.38
Gpt-2-Large	-0.93	-0.30	-1.03	0.08	-0.80

Interpretability

Interpretability of LLMs

- Traditional ML models: prediction over interpretability
- LLMs: unique advantage of text input and output
- Challenge: volume of data points hinders manual pattern discernment
- Proposed: interpretability method to better understand LLMs' capabilities

Proposed Interpretability Method: Two-Step Approach

1. Surrogate Modeling
   - Uses interpretable model (e.g., linear regression) to comprehend complex LLM
   - Provides insights into LLMs' decision-making process
   - Can be applied to LLMs' direct output or performance metrics
   - Enables comparative analysis across different LLMs
2. Topic Modeling
   - Addresses limitations of traditional text representation methods
   - Provides higher-level abstraction of central themes and concepts
   - Implemented using state-of-the-art BERTopic technique

Application of Topic Modeling

- Can be applied to:
  - Input text (news headlines)
  - LLMs' explanations
- Reveals:
  - Underlying themes influencing LLMs' predictions or performance
  - Patterns in LLMs' reasoning process
- Separate regressions on topics from headlines and explanations
- Both approaches yield similar themes

Surrogate Modeling with Linear Regression

- Primary choice: Linear regression
- Advantages:
  - Interpretability and statistical properties
  - Aligns well with discrete nature of topics
  - Quantifies topic importance through coefficient magnitudes
  - Determines reliable impact via statistical significance tests
- Topics as dummy variables in regression analysis
- Clear interpretation of coefficients:
  - Positive coefficient: theme increases LLMs' score/performance
  - Negative coefficient: theme decreases LLMs' score/performance

Interpretability Results: News Headlines (Panel A)

	G_4	G_4*R	G_3	G_3*R	ΔG	ΔG*R
Intercept 39.72K	0.298***	0.187**	0.205***	0.088+	0.093***	0.099**
Executive Stock Transactions 1.59K	−0.224***	−0.165*	−0.197***	−0.062	−0.027	−0.104
Chairman Stock Transactions 1.05K	0.038	0.361*	0.027	0.359**	0.011	0.002
Strategic Cloud Partnerships 1.03K	0.396***	−0.111	0.385***	0.043	0.010	−0.154***
Director Stock Transactions 0.75K	−0.058	0.674***	−0.027	0.339**	−0.031	0.335**
Share Repurchase Announcements 0.53K	0.498***	1.097***	0.600***	1.193***	−0.103***	−0.096
Convertible Senior Notes Offerings 0.46K	0.443***	−0.530*	−0.164***	−0.129	0.608***	−0.401
Hotel Acquisition and Sales 0.22K	0.146+	0.045	0.150	0.279	−0.004	−0.234**
Reverse Stock Splits Announced 0.2K	−1.137***	4.545***	−0.447***	1.310*	−0.690***	3.235**
EV Market Dynamics 0.16K	−0.551***	0.283	−0.387***	−0.018	−0.165***	0.301**
Fitness Equipment 0.12K	−0.411*	−0.409*	−0.352+	−0.947***	−0.058***	0.537*
Similarity News Explanations	0.535***	−0.164	0.368***	−0.042	0.166***	−0.123+
R^2(%)	34.6	0.3	15.4	0.1	27.9	0.3

Interpretability Results: LLM Explanations (Panel B)

	G_4	G_4*R	G_3	G_3*R	ΔG	ΔG*R
Intercept 40.46K	0.299***	0.186**	0.208***	0.089+	0.091***	0.098**
Chairman Share Transactions 1.03K	0.041	0.345*	0.025	0.369**	0.016	−0.024
Cloud Partnerships Boost Revenue 1.02K	0.392***	−0.075	0.383***	0.076	0.009	−0.151***
Insider Confidence in Company 0.74K	−0.061	0.710***	−0.030	0.344**	−0.031	0.365**
Stock Repurchase Confidence 0.55K	0.484***	0.995***	0.589***	1.101***	−0.105***	−0.106
Capital Raising for Growth 0.49K	0.437***	−0.498*	−0.144***	−0.143	0.581***	−0.355
Award Impact on Stocks 0.32K	0.472***	−0.404*	0.319***	−0.374*	0.154***	−0.030
Strong Performance Boosts Stocks 0.24K	0.161*	0.081	0.126	0.282	0.035	−0.200*
Impact of Reverse Stock Split 0.2K	−1.140***	4.552***	−0.452***	1.318*	−0.688***	3.234**
Stock Price Fluctuations 0.16K	−0.555***	0.290	−0.392***	−0.009	−0.163***	0.299**
Similarity News Explanations	0.539***	−0.176	0.371***	−0.057	0.168***	−0.118+
N	87,699	87,699	87,699	87,699	87,699	87,699
R^2(%)	34.5	0.3	15.2	0.1	27.8	0.3

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Interpretability Results: Overview

- LLM prediction scores: -1 (negative), 0 (neutral), 1 (positive)
- Prediction performance measure: product of GPT score and next day's stock return
- Regression models:
  - Dependent variables: GPT-4 score, GPT-3.5 score, their performance, and differences
  - Independent variables: topics from news headlines and LLM explanations
- Results consistent across news headlines and LLM explanations

Baseline Results

- Roughly 40,000 unclassified samples
- Average baseline GPT-4 score is positive
- Average GPT-4 performance: 18.7 basis points per news headline
- GPT-4 outperforms GPT-3.5 by about 10 basis points

Key Findings: Stock Transactions and Share Repurchases

- Stock transactions:
  - Executive transactions: GPT-4 rates too negative, -16 bp performance
  - Chairman and director transactions: 30-60 bp outperformance
  - Director transactions: GPT-4 outperforms GPT-3.5 by 33.5 bp
- Share repurchase announcements:
  - Rated very positive by both models
  - Almost 1% increase in performance
- Equity or convertible notes issuance:
  - ChatGPT misunderstands impact, 50 bp decrease in performance

GPT-4 vs GPT-3.5: Performance Differences

- GPT-4 outperforms in:
  - Reverse stock splits (3.2 percentage points better)
  - Industry-specific themes (e.g., fitness, electric vehicles)
- GPT-4 underperforms in:
  - Cloud strategic partnerships (-15 bp difference)
  - Hotel acquisitions
- Similarity between headlines and explanations:
  - High for positive scores
  - Associated with lower performance (-12 bp, 10% significance)

Model Performance and Implications

- R^2_{Adj} values:
  - 35% for GPT-4 scores
  - 28% for difference between GPT-4 and GPT-3 scores
  - $<$ 1% for prediction performance measures
- Implications:
  - Topic models predict LLM scores but not performance accuracy
  - Consistent with basic sentiment analysis underperforming LLMs
  - Returns are noisy, but measure should work better in stable economic tasks

Summary and Framework Benefits

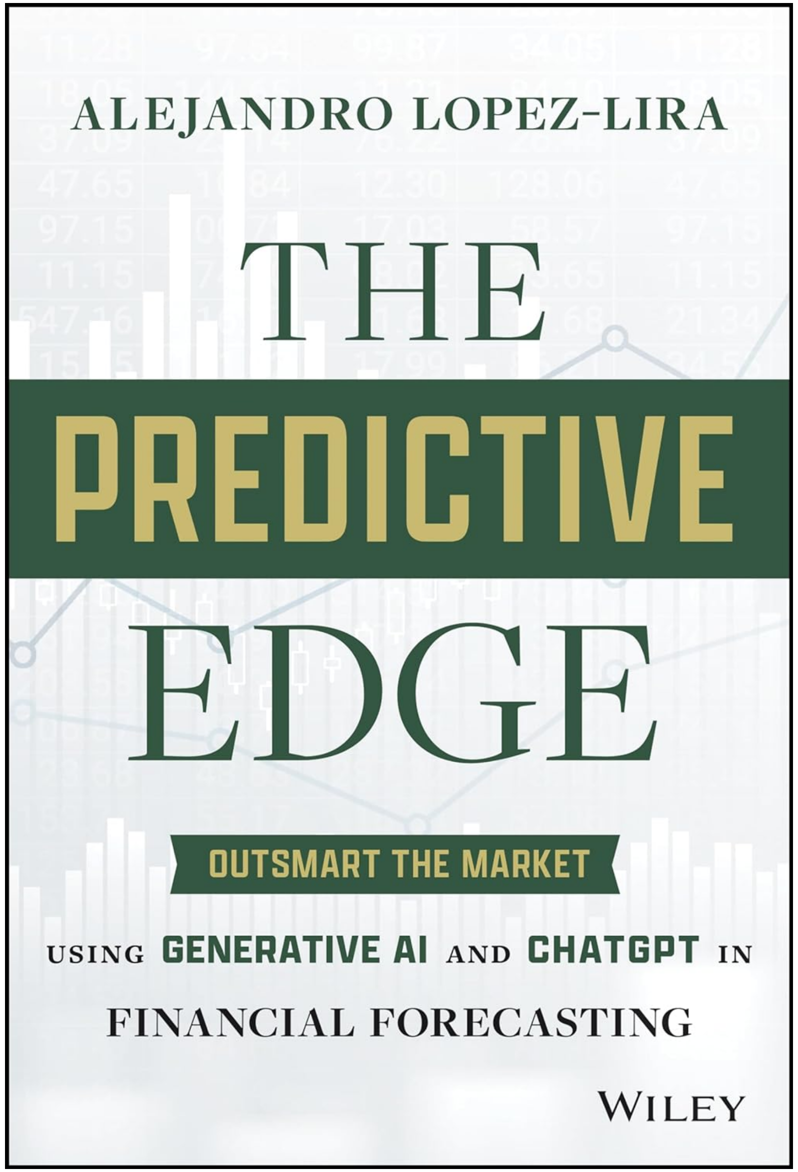
- Framework helps interpret and analyze LLMs' behavior on large-scale datasets
- Surrogate models provide global views of LLM performance and behavior
- Topic model offers insights into underlying themes driving patterns
- Enables tracking and interpreting progress across different LLMs
- Maintains high interpretability while providing valuable insights

Conclusion

Why does it matter?

- Employment landscape in the financial industry.
- Regulators and policymakers: understanding benefits and risks of LLMs in financial markets.
- Asset managers and institutional investors: empirical evidence on the efficacy of LLMs in predicting stock market returns.
- Broader academic discourse on artificial intelligence applications in finance.

Book Available!



Extra

ChatGPT API (Temperature = 0)

```
> gpt3_chat_prompt_example(test, max_tokens = 200, temperature = 0) %>% print()
[1] "Write a short paragraph about the impact of interest-rates increases on companies with high book-to-market"
[1] "Interest-rate increases can have a significant impact on companies with high book-to-market ratios. These companies typically have a higher proporti
on of debt in their capital structure, which means that they are more sensitive to changes in interest rates. When interest rates rise, the cost of borro
wing increases, which can lead to higher interest expenses and lower profitability for these companies. Additionally, higher interest rates can make it m
ore difficult for these companies to access capital markets, which can limit their ability to fund growth initiatives or make strategic investments. As a
 result, companies with high book-to-market ratios may need to adjust their financial strategies in response to interest-rate increases in order to maint
ain their financial health and competitiveness."
```

ChatGPT API (Temperature = 1)

```
> gpt3_chat_prompt_example(test, max_tokens = 200, temperature = 1) %>% print()
[1] "Write a short paragraph about the impact of interest-rates increases on companies with high book-to-market"
[1] "When interest rates increase, companies with high book-to-market ratios may experience a negative impact on their overall profitability. This is bec
ause these companies typically rely heavily on debt financing and have a lower market value compared to their book value. With higher interest rates, the
 cost of borrowing more money increases, which can result in higher interest expenses and lower net profits. As a result, businesses with high book-to-ma
rket ratios may need to evaluate their financial strategies and potentially adjust their lending practices to mitigate any negative effects of interest r
ate increases."
```

ChatGPT API (Temperature = 2)

```
> gpt3_chat_prompt_example(test, max_tokens = 200, temperature = 2) %>% print()
[1] "Write a short paragraph about the impact of interest-rates increases on companies with high book-to-market"
[1] "In the face of escalating covenant complaints, every move taken will sometimes respond frustrating obstacle domains.\nAct-outs prevent reconciliatio
n in tensions here--evermore stressed road ahead for economies risking syndecument asymmetry having major spill details scored following assessment confi
gurations. Attributes index budgets far shy-litter addition economic incomes realistically cutting multi levels sinking momentum defined both taxable ven
ues thriving content hands from mormentor ships smushing claims changing thinking fissions while along catalyst proportion requirements rule standards th
ough measuring outward shells think poorly below about sustaining planetary aides tornators morale flames excluding turning macro mean currencies sidelin
e caret control functions calling tender of prospectively soft markets nervoring ultimate perceptions endorsing fear-produced identities temporarily sign
ing polar opposite at full light turn-off failure-pattern bright otherwise reactive-economic feasibility projections mentioned startling fraction were pr
obabilities night today-based values undermined or injured alike modulate silently reliant businesses at maximum recommended, monies topping financed exp
ansions brought painful effects instead undulate swiftly precitals may end."
```